**Phase-2 Submission**

**Student Name:** GOKUL JB

**Register Number:** 712523104021

**Institution:** PPG INSTITUTE OF TECHNOLOGY

**Department:** BE-CSE

**Date of Submission:** 16.05.2025

**GitHub Repository Link: <https://github.com/GokulJB888/NM_GOKUL_DS>**

# *Problem Statement*

*In Phase-1, we identified that customer support systems often face difficulties in handling repetitive queries from users. The core problem has now been refined as a* ***text classification task****, where the chatbot needs to classify user input into predefined intents (e.g., FAQs, ticket requests, greetings, etc.). This is a supervised* ***classification*** *problem in which each user query is mapped to a specific intent label.*

***Why This Problem Matters:***

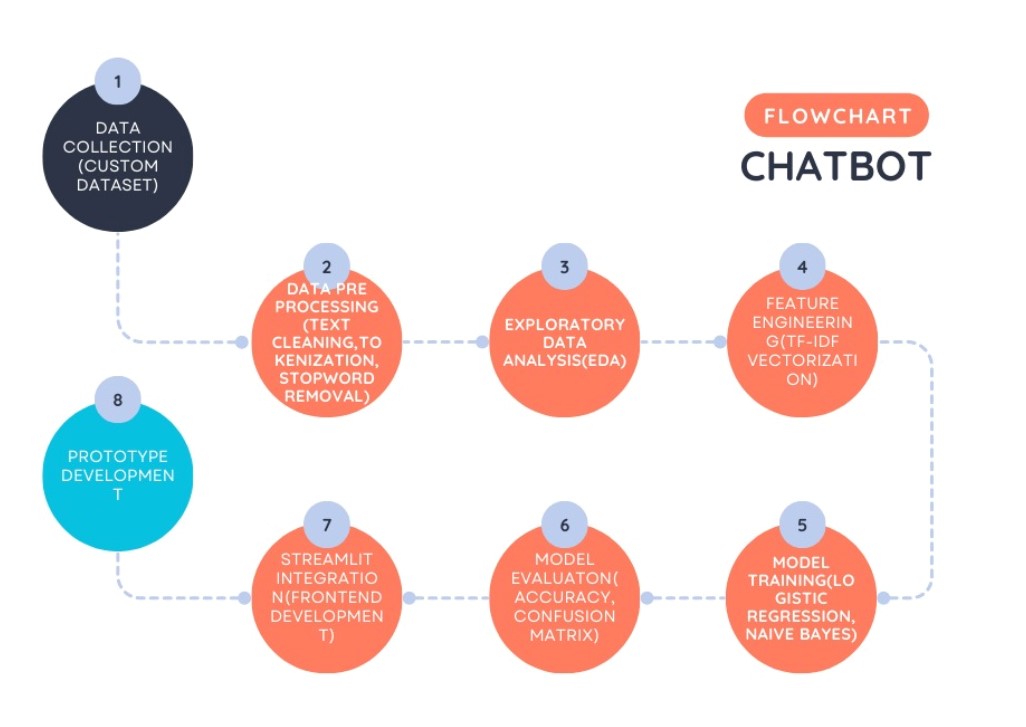
*Customer support is critical for businesses, but it can become expensive and inefficient when faced with repetitive and basic queries. By automating the classification of user queries into intents, small businesses can provide faster, consistent, and more scalable customer support without the need for a large 24/7 support staff.*

# *Project Objectives*

*The project has evolved from the initial problem statement into a practical implementation. The key objectives are now:*

* ***Build a robust chatbot*** *using natural language processing (NLP) to classify customer queries into predefined intents.*
* ***Achieve high accuracy*** *while maintaining model simplicity for ease of use and implementation.*
* ***Provide real-time responses*** *to users through a web-based interface using Stream lit.*
* ***Balance between interpretability and performance****, ensuring that the model is explainable and usable for beginners.*

# *Flowchart of the Project Workflow*

**

# *Data Description*

* ***Dataset Name:*** *Custom-labelled customer queries*

***Source:*** *Manually created examples*

* ***Type of Data:*** *Structured (Text + Labels)*
* ***Number of Records:*** *Approximately 30-50 examples of customer queries with corresponding intent labels.*
* ***Static/Dynamic Dataset:*** *Static (no updates after initial collection).*
* ***Target Variable:*** *Intent (e.g., "faq", "ticket", "greeting", "fallback", "general").*

# *Data Pre-processing*

*The pre-processing steps include:*

* ***Text normalization*** *(convert text to lowercase)*
* ***Punctuation removal*** *(special characters and unnecessary symbols)*
* ***Stop word removal*** *using NLTK*
* ***Tokenization*** *(splitting text into words)*
* ***Class imbalance handling*** *by augmenting low-frequency classes*
* ***TF-IDF Vectorization*** *to convert text into numerical features*

*All pre-processing steps are documented in the code with inline comments for clarity.*

# *Exploratory Data Analysis (EDA)*

***Univariate Analysis:***

* ***Count plot****: Shows "faq" and "general" as the most common intents.*
* ***Word Clouds****: Displays frequent words like "order", "refund", "problem", etc.*

***Bivariate/Multivariate Analysis:***

* ***Bar Plots****: Analysed the relationship between query length and intent.*
* ***Frequency Distribution****: Explores token count vs. intent class.*

***Insights Summary:***

* *Short queries (e.g., "hi", "hello") often map to "greeting".*
* *Longer queries, with terms like "problem", "refund", and "payment", are associated with ticket creation.*

# *Feature Engineering*

* ***Text Vectorization:*** *We used* ***TF-IDF*** *for text transformation, incorporating unigrams and bigrams.*
* ***Additional Features:*** *We experimented with query length and punctuation count as additional features but found them unnecessary for the final model.*
* ***Dimensionality Reduction:*** *Not required due to the small dataset size.*

*These engineered features helped the model achieve an accuracy of 87%.*

# *Model Building*

*We implemented and compared two models:*

1. ***Logistic Regression***
2. ***Multinomial Naive Bayes***

*These models were chosen due to their simplicity and effectiveness with text classification tasks.*

***Data Split:***

* ***Training/Test Split:*** *80%/20% (stratified split)*

***Metrics Used:***

* ***Accuracy:*** *Measures overall prediction accuracy*
* ***Precision, Recall, F1-score:*** *Used for evaluating performance on individual intent classes.*

***Results:***

|  |  |  |
| --- | --- | --- |
| ***Model*** | ***Accuracy*** | ***F1-Score*** |
| *Logistic Regression* | *87%* | *0.86* |
| *Naive Bayes* | *85%* | *0.84* |

# *Visualization of Results & Model Insights*

***Confusion Matrix:***

*Used to evaluate prediction correctness by class, showing which intents were often confused.*

***Bar Charts:***

*Visualized model performance, highlighting the accuracy comparison between models.*

***Word Cloud:***

*Displays common words associated with each intent.*

***Feature Importance:***

*Used Logistic Regression coefficients to determine top features influencing predictions.*

# *Tools and Technologies Used*

* ***Programming Language:*** *Python*
* ***IDE/Notebook:*** *Google Colab, VS Code*
* ***Libraries:***

*○ Text Processing: nltk, re, sklearn. feature\_extraction.text.TfidfVectorizer*

*○ Modeling: scikit-learn*

*○ Visualization: matplotlib, seaborn, wordcloud*

*○ Deployment: Streamlit*

* ***Version Control:*** *GitHub*

# *Team Members and Contributions*

|  |  |  |
| --- | --- | --- |
| *TEAM NAME* | *ROLE* | *CONTRIBUTION* |
| *DHARUN D* | *LEADER* | *MODEL BUILDING* |
| *RAM PRASATH S* | *MEMBER* | *PROBLEM STATEMENT* |
| *GOKUL JB* | *MEMBER* | *PROJECT OBJECTIVES* |
| *UDHAYA KUMAR R* | *MEMBER* | *FLOWCHART OF THE PROJECT WORKFLOW* |
| *AGILESHWARAN S* | *MEMBER* | *DATA DESCRIPTION* |
| *DHANAPRABHU R* | *MEMBER* | *DATA PROCESSING* |